INF5830 – 2017 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning, Lecture 5, 19.9

Today

- Estimation
  - general case
  - for a proportion
- Comparing two independent
  - populations
  - proportions
- Paired data
  - Sign test, McNemara's test
  - Paired t-test

Last week: Why statistics in evaluation?

- Task1: Completed last week
  - You know the best classifier on a task has 0.8 (80%) accuracy (baseline).
  - You have made a classifier which classify 85 items correctly on a test set of 100 items.
  - Can you conclude your classifier is better than the baseline?
- Task 2: Remains
  - You have made a classifier. You test it on 500 items. It classifies 375 correctly.
  - What is the accuracy of your classifier?

Why? (this week)

- Task 3:
  - You have two different classifiers, one with accuracy 0.89 and one with accuracy 0.91on 1000 test items.
  - Can you conclude that one is better than the other?
- Task 4:
  - The two classifiers from task 3 agree on 870 items.
  - One is doing better on 20 items, the other is doing better on 40 items.
  - Can we draw conclusions from this?

Estimation

The task

Last week:

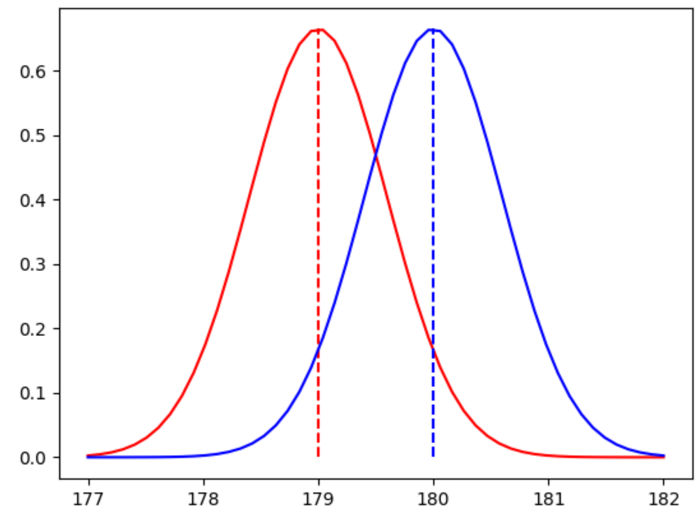
- Given:
  - A population P with known distribution (B(n,p) or N(μ, σ))
  - A SRS X
- Q: How probable is it that X could have been drawn from P?
  - What is the p-value?

Now:

- Given:
  - A SRS X, with a mean $\bar{X}$
  - A p-value: p
- Q: From which populations P could X have been drawn with prob. p?
- Q: In part. what is the mean, μ, for such a P?

Example

- Assume a population P2 and an SRS of 100 indiv.s from P2 with $\bar{x} = 179$
- What is μ for P2?
- Goal: find an e such that $P(179 - e < \mu < 179 + e) < p$ for some level p, e.g. 0.05
- Observe that $P(179 - e < \mu < 179 + e)$ $= P(\mu - e < 179 < \mu + e)$
- If we know the standard deviation, …



- … we could have calculate this simlarly to before
  - Find the z-value z corresp. to p
  - $e = z\frac{\sigma}{\sqrt{n}}\left(= z\frac{\sigma}{\sqrt{100}}\right)$

Estimation

- How to estimate the true mean μ if the standard deviation σ of the population is unknown?
- All we have is a sample $X = \{x_1, x_2, \ldots, x_n\}$
- The sample mean $\bar{x}$ is still the best point estimate of the pop. mean μ

Estimation

- To approximate the true standard deviation:
- We use the standard deviation of the sample X,
  - $s^2 = ((x1 - \overline{x})^2 + (x2 - \overline{x})^2 + \ldots + (xn - \overline{x})^2)/(n - 1)$
  - Observe (n-1) and not n
  - The (n-1) is to compensate for using $\overline{x}$ instead of μ in the formula
  - S is sometimes called standard error

s is a random variable (like $\overline{X}$) over all SRS of size n
s is an unbiased estimator for σ: E(s)= σ

Estimation

- In addition we do not use the standard Z-distribution but the t-distribution for n-1.
- Then the level C confidence interval for μ is
  - $[\bar{x} - e, \ \bar{x} + e]$
  - Where $$e = t^* \frac{s}{\sqrt{n}}$$
  - and t^* is the value from the $t(n-1)$ density curve for C

The t-distribution is similar to the z-distribution for large n.
But t^* is larger than z for same p when n is small

Example

- Assume we do not know the st.dev. 18 ys old men from Finmark
- Pick a random sample of 9 men:
  - $\bar{x} = 177$, s = 5
- Estimate the average height for this population
  - Choose confidence level 0.95

Table, or

```
In   [78]: stats.t.ppf(.025,8)
Out[79]: -2.306004135033709
```

Since n=9, we use the t(8)-density curve Jargon: There are 8 degrees of freedom

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 177 \pm 2.306 \frac{5}{\sqrt{9}} = 177 \pm 3.843$$

- The 95% confidence interval for μ: [173.1, 180.9]
- Exact for normal distribution
- Approximation for large n otherwise

What would be different if we used normal distribution?

Estimation with proportion

- Task 2:
  - You have made a classifier. You test it on 500 items. It classifies 375 correctly.
  - What is the accuracy of your classifier?

Proportion

- The best estimate we have for p is $\hat{p} = \frac{375}{500} = 0.75$
- The best estimate we have for the standard deviation is
  $$\mathrm{SE}(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \left(= \sqrt{\frac{0.75(1-0.75)}{500}} = 0.01934 \right)$$

Example

- Using normal distribution approximation:
  - In [284]: stats.norm.ppf([0.025, 0.975],0.75, np.sqrt(0.75*0.25/500))
  - Out[284]: array([0.71204546, 0.78795454])
- Using binomial distribution:
  - In [288]: stats.binom.ppf([0.025, 0.975],500, 0.75)/500
  - Out[288]: array([0.712, 0.788])

Comparing populations/proportions

Why? (this week)

- Task 3:
  - You have two different classifiers, one with accuracy 0.89 and one with accuracy 0.91on 1000 test items.
  - Can you conclude that one is better than the other?

The general case

- You have two populations P1 and P2, say Swedish and Italian 18 ys old men.
- You want to compare a variable between the populations, say height:
  - Either one-sided: Are men in P1 taller than men in P2?
  - Or two-sided: Have men in P1 different average height than men in P2?
- You don't know the true mean or st.dev of P1 nor of P2.

T-test for differences

- Procedure:
  - Draw a SRS from P1 with
    - n_1 individuals
    - mean $\bar{x}_1$
    - sample s.d. s_1
  - Similarly for P2
- Calculate the t-score
- Find *p* from the t(m)-density curve
  - What is *m?*
    - $m = \min(n_1, n_2)$
    - Scikit uses $m = n_1 + n_2$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{s_1}^2}{n_1} - \frac{{s_2}^2}{n_2}}}$$

Comparing proportions

- Task: compare two proportions:
  - Could they be SRS from the same population?
- Sample 1: n items, k successes
  $\hat{p} = {}^{k}/_{n}$ and $s^2 = \hat{p}(1-\hat{p})$
- Sample2: *m* items, *h* successes
  $\widehat{p_2} = {}^{h}/_{m}$ and ${s_2}^2 = \widehat{p_2}(1-\widehat{p_2})$
- Calculate the score $z = \frac{\hat{p}-\widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})/_{n}+\widehat{p_2}(1-\widehat{p_2})/_{m}}}$,
- Use the z-density curve to find p-value

Example

- Two classifiers:
- C1 has accuracy 0.876 on 500 items
- C2 has accuracy 0.896 on 500 items

Comparing accuracy

- $z = \frac{\hat{p} - \widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})/_n + \widehat{p_2}(1-\widehat{p_2})/_m}}$, use the z-density curve
- $\hat{p} = 0.896$, $\widehat{p_2} = 0.876$, n=m=500
- $z = \frac{0.896 - 0.876}{\sqrt{0.896\,(1-0.896)/_{500} + 0.876\,(1-0.876)/_{500}}}$=0.9995
- p-value = 0.16

Paired data

Example contd.

		Classifier 2	
		correct	incorrect
Classifier 1	correct	435	13
	incorrect	3	49

- It turns out the two classifiers were tested on the same 500 items
- We can record for each item whether each classifier is correct or not
- We expect them to be equally good

		Classifier 2	
		correct	incorrect
Classifier 1	correct	435	13
	incorrect	3	49

- We can focus on the items were they disagree.
- We expect (C1-correct & C2-incorrect) and (C2-correct and C1-incorrect) to be equally likely
- How unlikely is it that from the 16 items where they disagree, C1 wins 3 (or fewer) times?

Sign test


```
                              ||     Classifier 2     ||
                              ||----------------------||
                              || correct | incorrect  ||
======================================================||
              | correct       ||   435   |     13     ||
Classifier 1  |---------------||---------|------------||
              | incorrect     ||    3    |     49     ||
======================================================||
In [468]: stats.binom.cdf(3,16,0.5)*2
Out[468]: 0.021270751953125
```

If it is two-sided

Sign test

- This test is called the sign-test
- Can be used to numerical data
- When used on Boolean values, {0, 1}, often called McNemar's test

Paired t-test

- $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$
- x_k and y_k are observations of the same individual,
  - e.g. before and after treatment
  - Classifier_x and classifier_y's result on item *k*
- Let $z_k = y_k - x_k$ and perform a one-sample t-test on $Z = \{z_1, \ldots, z_n\}$, comparing to $\mu = 0$.
- This test can also be applied to proportions when the sample is big.

Paired t-test example


```
                          ||     Classifier 2     ||
                          ||----------------------||
                          || correct | incorrect  ||
==================================================||
              | correct   ||  435    |    13      ||
Classifier 1  |-----------||---------|------------||
              | incorrect ||   3     |    49      ||
==================================================||
Y = [1]*(435+13) + [0]*(3+49)
X = [1]*435 + [0]*13 + [1]*3 + 49*[0]
Z = [y - x for (y,x) in zip(Y,X)]
```

In SciPy

- In [479]: Y = [1]*(435+13) + [0]*(3+49)
- ...: X = [1]*435 + [0]*13 + [1]*3 + 49*[0]
- ...: Z = [y - x for (y,x) in zip(Y,X)]

- stats.ttest_rel(Y,X)
- Out[480]: Ttest_relResult(statistic=2.5132559949600304, pvalue=0.012276223171277646)

- In [4]: stats.ttest_1samp(Z,0)
- Out[4]: Ttest_1sampResult(statistic=2.5132559949600304, pvalue=0.012276223171277646)

Paired data

- The sign test is non-parametric
- The paired t-test assumes (nearly) normal distributions (OK for proportions with large *n* (>25))
- The paired t-test yields better numbers.

What we have done

1. How likely is a sample given a known population?
   1. How good is a classifier compared to a baseline?
2. From a sample, estimate an interval for the true mean value.
3. Comparing two independently drawn samples.
4. Comparing paired data.